Human-Assisted Agent for Sequential Decision-Making

Junzhe Zhang
Purdue University
zhang745@purdue.edu

Elias Bareinboim
Purdue University
eb@purdue.edu

ABSTRACT

Many real-world sequential decision-making settings involve human agents who make decisions having access to information that may be unavailable to artificial, automated systems. Standard reinforcement learning (RL) methods do not usually model human behavior and, almost invariably, attempt to fully replace the human component with a fully automated agent, which we call human-replacing agent (HR agent). In this paper, we propose the human-assisted agent (HA agent), which takes the natural agent's decision x' as a suggestion to decide a new treatment x maximizing the target outcome – that is, what the target outcome would have been, counterfactually, had the treatment been $X = x$, given that in fact $X = x'$. A proper causal language is required to support this mode of reasoning and allow valid counterfactual inference. We model this decision-making setting using the language of structural causal models [Pearl, 2000] and formulate the HA agent in counterfactual semantics. We identify a general class of SCMs where an HA agent can leverage MDP machinery and find a strategy that provably dominates previously known strategies, including HR-based ones. When the (counterfactual) Markov property does not hold, we equip the HA agent with an augmented POMDP capability. We further characterize the conditions under which natural agents' decisions do not offer valuable information to the task, which means that the performance of HA and HR agents coincide and autonomy can be reached.

CCS Concepts

•Computing methodologies → Causal reasoning and diagnostics;

Keywords

Causal Inference; Reinforcement Learning; Markov Decision Process

1. INTRODUCTION

We study decision-making settings where a human being (also called natural agent) makes decisions in an environment and can be influenced by multiple exogenous factors. These factors may have an effect on the agent's decisions consciously or subconsciously and are not necessarily recorded or accessible by other agents other than the original decision-maker. For example, consider a physician who decides the dose of a drug in different time slots (treatment) according to his perception of the patient's conditions (e.g., balance, dilation of the pupil, mood). With the accumulation of data and processing power, hospitals are increasingly interested in deploying computer agents to perform automatic diagnosis and planning of patients' treatments. One of the reasons is that it is not unreasonable to expect that physicians may be bound to find sub-optimal policies given their inability of quantifying uncertainty and processing huge amounts of data. On the other hand, however, automated systems may not have the necessary communication skills to fully access patients' conditions, thus being potentially unable to find an optimal policy. The natural question that arises, in this case, is whether it is justifiable to design an agent to utilize the benefit of automation and data analysis while leveraging the physician's experiences; if the answer is positive, what would be the principles that should drive the construction of such a system.

We propose the design of a human-assisted agent (HA agent) that takes the physician's decision as a suggestion and decides a possibly new course of treatment. The challenge of designing such an agent arises dues to the counterfactual nature of the problem, that is, what chances of recovery would a patient have had she taken a different treatment $X = x$, given that in fact, she is under treatment $X = x'$ [8, Ch. 1]. This counterfactual quantity seems to defy empirical experiences because we can never rerun history and administer a different level of treatment x for those who already received it at level x'. Performing this type of reasoning relies on the introduction of causal inference machinery [8, Sec. 7.1]. We will use Structure Causal Models (SCMs) as the basis of our analysis so as to be able to reason with counterfactual statements and perform inferences of key concepts, including counterfactual independences, conditional interventions, and expected outcomes.

Connections between the causal inference and reinforcement learning (RL) were first established in [1]. In this setting, Bareinboim et al. implicitly described an HA agent for Multi-Armed Bandits (MABs) where unobserved confounders (UCs) affect the agent's decision process It's well understood that MABs are a rather simplistic model where rewards and actions at different rounds are assumed to be mutually independent. Our approach here focuses on a more general decision-making setting where the action not only affects the immediate reward, but also the future state of the agent. Finding the optimal policy in such an environment requires non-trivial analysis of independence relationships among counterfactual variables, which hold by default in [1]. We will show that subtle mistakes could occur in designing inference algorithms for the HA agent when the recognition

of these independence relations are not explicit, which in turn would translate into a lack of convergence.

Counterfactual inference has been studied and applied in the context of RL under the rubrics of off-policy learning. [9, 5] applied the inverse propensity score weighting to estimate the effect of a new policy using samples collected by the agent running a different behavioral policy (the physician's policy in the medical treatment example). Most off-policy learning methods assume that the behavioral policy and the target policy share the same state-action spaces, which in practice, however, is not rarely violated. For instance, in the medical example, the physician could observe states unobserved to the learning agent (e.g., balance, mood). Also, current RL methods focus almost exclusively on human-replacing agents (HR agents) that are designed to substitute the existent natural agent from the environment. From a causal perspective, off-policy methods estimate the average effect of a treatment x on the general population, instead of the effect in the specific population that is currently under treatment x'; the latter might possess distinct needs and dispositions that make them react differently to a different treatment x than a randomly selected subject would.

In this paper, we model this decision setting with SCMs and analyze HA agents with formal counterfactual language. Specifically, our contributions are as follow:

1. We first show that HR agents are not guaranteed to behave optimally in an environment where natural agents are present (e.g., social and medical settings), and decisions are driven by unobserved confounders.

2. We identify a class of models where the optimal policy of the HA agent can be obtained through a simple modification of MDP algorithms, contrasting with standard HR agents.

3. For systems not contained in the above class (i.e., when the counterfactual Markov property does not hold, to be defined), we propose a modified POMDP planning algorithm to find the optimal policy of HA agents.

4. We prove a general condition where it is safe to replace the natural agent with a standard agent. The proof confirms the intuition that RL algorithms do not need input from the natural agent only when it has superior capabilities for observing all the latent states.

2. PRELIMINARIES AND NOTATIONS

In this section, we introduce the basic notations and definitions used throughout the paper. We will consistently use the abbreviation $P(x)$ for the probabilities $P(X = x)$, where x is an arbitrary value.

We describe the environment using the Structural Causal Models (SCMs) defined in [8, pp. 203-205]. SCMs gives formal meaning for fundamental concepts, including confounding, observational and experimental distributions, and counterfactuals. We define SCMs in the sequel.

DEFINITION 1. *(SCM [8]). A structural causal model (SCM) M is a 4-tuple $\langle U, V, F, P(u)\rangle$ where:*

1. *U is a set of exogenous (unobserved) variables, that are determined by factors outside of the model,*
2. *V is a set $\{V_1, V_2, \ldots, V_n\}$ of endogenous (observed) variables that are determined by variables within the model (i.e., by the variables in $U \cup V$),*
3. *F is a set of function $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from the respective domain of $U_i \cup PA_i$ to V_i, where $U_i \subseteq U$ and $PA_i \subseteq V \backslash V_i$ and the entire set F forms a mapping from U to V. In other words, each f_i, $v_i \leftarrow f_i(pa_i, u_i), i = 1, \ldots, n$, assigns a value to V_i that depends on the valus of the select set of variables,*
4. *$P(u)$ is a probability distribution over the exogenous.*

Each SCM M is associated with a directed acyclic graph (DAG) G, where solid nodes correspond to endogenous variables V, empty nodes correspond to exogenous variables U, and edges represent functional relationships (see Fig. 1).

The mathematical operator $do(X = \pi(w))$, defined in [8], denotes an intervention where the values of X are set according to an arbitrary function $\pi(w)$, regardless of how the values of X are ordinarily determined in the model[1]. We use a causal effect $P(Y = y|do(X = \pi(w)))$ to represent the response of a variable Y to the intervention $do(X = \pi(w))$. This causal effect is sometimes denoted by a counterfactual quantity $P(Y_{X=\pi(w)} = y)$. The formalism of SCMs allows a probabilistic measure over counterfactual variables, i.e., $P(Y_{X=\pi(w)} = y) = \sum_{\{u \in E\}} P(u)$, where E is a set of realizations of U compatible with $Y = y$ in the post-interventional model under $do(X = \pi(w))$ [8]. If $\pi(w) = x$ where x is an arbitrary constant, the intervention $do(X = x)$ is called atomic. We will use the abbreviation $P(y_x)$ for distributions $P(Y_{X=x} = y)$.

In this paper, we consider a sequential decision problem, where at time $t = 1, 2, \ldots$, the agent observes the state $S^{(t)} = s^{(t)}$, performs an action $do(X^{(t)} = x^{(t)})$, receives an reward $Y^{(t)} = y^{(t)}$, and moves to the next state $S^{(t+1)} = s^{(t+1)}$. We will consistently use the abbreviation $x^{([1,t])}$ for a sequence $\{x^{(1)}, x^{(2)}, \ldots, x^{(t-1)}\}$. Let $h^{(t)}$ be the observable history up to time t, $H^{(t)}$ be the set of all possible histories up to time t, and let S, X, Y be finite domains for (respectively) states, actions and rewards. In causal semantics, $h^{(t)} = \{s^{(0)}, s^{(1)}_{x^{(1)}}, \ldots, s^{(t)}_{x^{([1,t-1])}}\}$, where $s^{(t)}_{x^{([1,t-1]}}$ represents the event $S^{(t)} = s^{(t)}$ after past actions $x^{([1,t-1])}$. Define a decision rule at time t to be a distribution function $\pi^{(t)} := H^{(t)} \times X \to [0, 1]$. A policy Π for an agent is a sequence of policies, that is, $\Pi = \pi^{([1,t])}$. A policy Π is called a stationary Markov policy if $\pi^{(t)} = \pi$ at any time t, where $\pi := S \times X \to [0, 1]$. We use the decision rule π in short for such a stationary Markov policy. We define the expected discounted cumulative rewards starting from state $s^{(1)}$ under the policy Π by

$$V^{\Pi}(s^{(1)}) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t Y^{(t)}_{X^{([1,t-1])}=\pi^{([1,t-1])}}],$$

where the subscript $X^{([1,t-1])} = \pi^{([1,t-1])}$ stands for the intervention $do(X^{(1)} = \pi^{(1)}, \ldots, X^{(t-1)} = \pi^{(t-1)})$, and $\gamma \in [0, 1)$ is a discount factor. The goal is to find an optimal policy Π^* maximizing the expected discounted cumulative rewards.

We use the formalism of finite Markov decision processes as the main tool to solve for the optimal policy.

[1] *Ordinarily* here is also called "naturally" in the causal inference literature, and represents a "behavioral" policy.

Figure 1: (a) The SCM for the medical treatment example described in Sec. 3, denoted by MDPUC. (b) The SCM with which the HA agent cannot be cast to a MDP, due to the confounding edge $S^{(t)}, X^{(t)}$.

Definition 2. *(MDP [2]). A finite Markov decision process (MDP) is a 5-tuple $\langle S, X, Y, T, R\rangle$ in which S is a finite set of states; X a finite set of actions; Y a set of rewards; T a transition distribution defined as $T : S \times X \times S \to [0, 1]$; and R a reward distribution defined as $R : S\times X\times Y \to [0, 1]$.*

We can model a decision setting as a MDP if it is Markov, i.e, the result of an action does not depend on the previous actions and observations (history), but only depends on the current observation. A number of efficient algorithms have been studied to solve MDPs, including value iteration [11] for offline planning problems and MORMAX [12] for online learning problems.

A system might be no longer Markov if unobserved elements of the state affect the next state. To find an optimal solution in this case, one popular approach is to model it as a partially observable MDP (POMDP), where the agent does not access the state $S^{(t)} = s^{(t)}$ but a noisy observation $O^{(t)} = o^{(t)}$.

Definition 3. *(POMDP [4]). A finite Markov decision process (MDP) is a 7-tuple $\langle S, X, Y, O, T, R, \Omega\rangle$ in which S, X, Y, T, R are the same defined in Def. 2; O a finite set of observations; and Ω an observation function defined as $\Omega : S \times X \times O \to [0, 1]$.*

Planning on a POMDP is equivalent to solving for a MDP with a continous state space, which is intractible in general [6]. Approximation planning algorithms, e.g., incremental pruning [3], which exploits the structure of the optimal POMDP policy, have been extensively studied and shown to be efficient, whereas online learning algorithms were not well understood until recently [10].

3. A MOTIVATING EXAMPLE

We start this section by revisiting the medical treatment example mentioned in the introduction. A physician treats a patient who visits the hospital regularly to maintain his long term health condition. At the t-th visit, the physician measures the patient's corticosteroid level $S^{(t)} = s^{(t)}, s^{(t)} \in \{0, 1\}$, where 0 stands for a low and 1 for a high level of corticosteroid. She then decides a treatment $X^{(t)} = x'^{(t)}, x'^{(t)} \in \{1, 0\}$ (1 for to give the drug, 0 for not to), and then measures an overall health score $Y^{(t)} = y^{(t)}, y^{(t)} \in \{1, 0\}$ (i.e., "healthy" and "not healthy"). In reality, the patient's health score $Y^{(t)}$ is also affected by a pair of factors $U^{(t)} = \{M^{(t)}, E^{(t)}\}$ where $M^{(t)} = m^{(t)}$ stands for patient's psychological status (0 for positive, 1 for negative) and $E^{(t)} = e^{(t)}$ stands for his socioeconomic status (0 for wealthy, 1 for poor). We model the patient's longterm health condition by the discounted cumulative reward with $\gamma = 0.99$. The physician follows a stationary Markov policy π^{ndt} where $X^{(t)} \leftarrow \pi^{ndt}(s^{(t)}, m^{(t)}, e^{(t)})$. Despite affecting the physician's decision, $m^{(t)}$ and $e^{(t)}$ are not recorded in the hospital's database due to privacy concerns. The full parametrizations of this structural model with the reward function $P(Y^{(t)} = 1 \mid s^{(t)}, m^{(t)}, e^{(t)}, x^{(t)})$ and transition function $P(S^{(t+1)} = 0 \mid s^{(t)}, x^{(t)})$ is provided in Sec. 6.

Unsatisfied with the condition of the patient, the hospital decides to replace the physician with a computer agent (HR agent). Fig. 1(a) shows the SCM of the system from an agent's perspective for $t = 3$, where $U^{(t)}$ is unobserved. At each visit t, the agent decides a treatment $X^{(t)} = x^{(t)}$ following a policy $\pi^{(t)}$ given history $h^{(t)}$. Due to the existence of unobserved variables $U^{(t)}$, we can cast this problem to a POMDP and solve for the optimal policy. However, with careful examinations, we find this system is indeed Markov, i.e.,

Theorem 1. *Consider the SCM described in Fig. 1 (a), starting from an arbitrary state $s^{(1)} \in S$, it satisfies following properties:*

$$P\left(s^{(t+1)}_{x^{([1,t])}} \mid h^{(t)}\right) = P\left(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}\right) \quad (1)$$

$$P\left(y^{(t)}_{x^{([1,t])}} | h^{(t)}\right) = P\left(y^{(t)}_{x^{(t)}} | s^{(t)}\right). \quad (2)$$

Proof. Introducing interventions $do(X^{([1,t])} = x^{([1,t])})$ is equivalent to removing all incoming edges of $X^{([1,t])}$. In the post-interventional SCM, $S^{(t)}$ becomes the only variable between $S^{(t+1)}, Y^{(t)}$ and past states and actions. Therefore, $S^{(t+1)}, Y^{(t)}$ are d-separated from the history given $S^{(t)}$, which proves Eqn. 1 and 2. □

Thm. 1 says that for the medical treatment example, the system is Markov for a HR agent. We thus can solve for the optimal policy for the HR agent with a standard MDP planning algorithm, and the optimal policy must be a stationary Markov policy π where $\pi := S \to X$.

We label the optimal policy learned by a MDP planning algorithm as *mdp*, and the one learned by a POMDP planning algorithm as *pomdp*. We also include three baseline policies for comparison: 1. the physician's current policy without any intervention (called *ndt*), 2. a policy picking the treatment at random (*random*), and 3. the true optimal policy learned by an agent who can access all unobserved information (*opt*). Fig. 2 shows the cumulative reward and

average reward per episode of these experiments. Somewhat surprisingly, none of the algorithms is able to learn a reasonable policy – also, the results coincide with the random policy. Moreover, the *ndt* policy performs worse than random guessing.

The experimental results suggest that the HR agent is not able to converge to some acceptable policy. This raise the question of whether a HA agent which, instead of replacing the physician, takes the physician's decision as a suggestion could perform better in this environment. At the t-th visit, a HA agent observes not only the current state $S^{(t)} = s'^{(t)}$, but the physician's decision $X^{(t)} = x'^{(t)}$ as well (but never $U^{(t)}$), and finally picks a new treatment $X^{(t)} = x^{(t)}$. Let $h'^{(t)}$ be the observable history for a HA agent by the time t wehere $h'^{(t)} = h^{(t)} \cup \{x'^{(1)}, x'^{(2)}_{x^{(1)}}, \ldots, x'^{(t)}_{x^{([1,t-1])}}\}$ and let $H'^{(t)}$ denote all possible $h'^{(t)}$. A HA agent follows a policy Π where $\Pi = \pi^{([1,t]}$ and $\pi^{(t)} := H'^{(t)} \times X \to [0, 1]$.

Two natural questions arise at this point: 1. how HA agent can systematically find an optimal policy; 2. in which cases an HA agent outperforms an HR agent and a human agent. Our goal in the remaining of the paper is to answer these two questions.

4. HUMAN-ASSISTED AGENT AS A MDP

In this section, we study whether (and, if so, how) an HA agent can be modelled using standard MDP machinery. Our goal is to construct a transformation of HA agents to MDPs so as to apply the corresponding MDP algorithms. The goal then is to analyze the behavior of such mapping in terms of optimality.

To cast an HA agent to an MDP agent, we first need to show that the corresponding system is Markov, that is,[2]

$$P\left(s^{(t+1)}_{x^{([1,t])}} \mid h'^{(t)}\right) = P\left(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right) \quad (3)$$

$$P\left(y^{(t)}_{x^{([1,t])}} \mid h'^{(t)}\right) = P\left(y^{(t)}_{x^{(t)}} \mid s^{(t)}, x'^{(t)}\right). \quad (4)$$

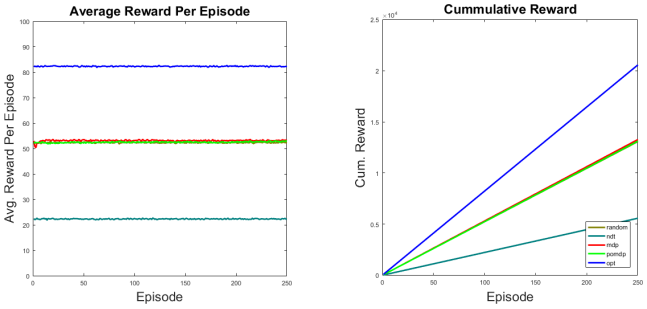
We first focus on the system described in the previous section, which is called the MDPUC due to the presence of unobserved confounders (UCs) $U(t)$. Let S, X, Y, U denote by domains of the state $S^{(t)}$, action $X^{(t)}$, reward $Y^{(t)}$, and the exogenous variables $U^{(t)}$. We note that a MDPUC running with a HR agent is a MDP.

THEOREM 2. *(MDPUC Markovianity) For the MDPUC starting from an arbitrary state $s^{(1)} \in S$, Eqn. 3 and 4 hold.*

We will show next the proof of the above statement. The proof builds on the graphoid axioms [7], the exclusion restrictions rule of SCMs [8, pp. 232], and three axioms of structural counterfactuals: composition, effectiveness and reversibility [8, Sec. 7.3.1]. We also build on confounded components (C-components), a useful concept for operating SCMs with unobserved confounders ([13]).

DEFINITION 4. *(C-component [13]) Let G be a causal diagram such that a subset of its bidirectional dashed edges (connected by exogenous variables) forms a spanning tree over all endogenous variables in G. Then G is a C-component.*

[2]The variable X taking different values before and after the conditioning bar (i.e., x' and x) syntactically exhibits the counterfactual nature of the problem.



(a) (b)

Figure 2: Average reward per episode plot (left) and cumulative reward plot (right) for medical treatment example.

For example, the SCM described in Fig. 1 contains 3 C-component: $\{S^{(1)}\}, \{X^{(1)}, Y^{(1)}, S^{(2)}\}$ and $\{X^{(2)}, Y^{(2)}, S^{(3)}\}$ since the variables in each of these components are connected through bidirected arrows. Every SCM can be uniquely partitioned into a set of C-components $C^{(1)}, C^{(2)}, \ldots, C^{(K)}$ where $K < \infty$. Let $U^{(k)}$ denote by the set of exogenous variables associated with a C-component $C^{(k)}$. The SCM assumes that $\{U^{(1)}, U^{(2)}, \ldots, U^{(K)}\}$ are mutually independent in the joint probability $P(u)$. This assumption can be translated into a probabilistic decomposition formula in terms of counterfactual statements.

LEMMA 1. *Given a SCM $M\langle U, V, F, P(u)\rangle$ with N endogenous variables and K C-components. Let PA_X be parents of a variable X, and let $V_{pa} = \{V^{(1)}_{pa_{V^{(1)}}}, \ldots, V^{(N)}_{pa_{V^{(N)}}}\}$. The joint distribution $P(v_{pa})$ factorizes according to the product:*

$$P(v_{pa}) = \prod_{k=1}^{K} P(c^{(k)}_{pa}), \quad (5)$$

where $C^{(k)}_{pa}$ denote endogenous variables with parents' value fixed in C-component k by $C^{(k)}_{pa} = \{V^{(1)}_{pa_{V^{(1)}}}, \ldots, V^{(N^{(k)})}_{pa_{V^{(N^{(k)})}}}\}$.

PROOF. Let $U^{(k)}, C^{(k)}$ be, respectively, exogenous and endogenous variables of C-component k. A $V^{(i)}_{pa_{V^{(i)}}}$ in $C^{(k)}_{pa}$ is decided by the function $V^{(i)} = f(pa_{V^{(i)}}, u^{(k)})$ with its parents' value fixed at $PA_{V^{(i)}} = pa_{V^{(i)}}$. We can thus write $P(c^{(k)}_{pa})$ as:

$$P(c^{(k)}_{pa}) = \sum_{U^{(k)}} \prod_{i=1}^{N^{(k)}} I\{v^{(i)} = f(pa_{V^{(i)}}, u^{(k)})\} P(u^{(k)}) \quad (6)$$

Since U is partitioned into mutually independent subsets $\{U^{(1)}, U^{(2)}, \ldots, U^{(N)}\}$, $P(v_{pa})$ can be written as

$$P(v_{pa}) = \sum_{U} \prod_{j=1}^{K} \prod_{n=1}^{N^{(j)}} I\{v^{(n)} = f(pa_{V^{(n)}}, u)\} P(u^{(j)}).$$

By moving out $U^{(k)}$ and $C^{(k)}$, this becomes:

$$P(v_{pa}) = \underbrace{\sum_{U^{(k)}} \prod_{i=1}^{N^{(k)}} I\{v^{(i)} = f(pa_{V^{(i)}}, u^{(k)})\} P(u^{(k)})}_{\text{Part 1}} \cdot \underbrace{\sum_{U \setminus U^{(i)}} \prod_{j=1, j\neq k}^{K} \prod_{n=1}^{N^{(j)}} I\{v^{(n)} = f(pa^{(n)}, u^{(n)})\} P(u^{(j)})}_{\text{Part 2}}$$

Part 1 is exactly $P(s^{(j)}_{pa})$ defined in Eqn. 6. By applying the same procedure for remaining $K-1$ C-components, we obtain Eqn. 5. □

Lem. 1 implies a stronger result than the independence restrictions rule [8, Sec. 7.3], since it shows that all endogenous variables with their parents fixed are independent of variables with parents fixed in other C-components.

Proof. (Proof of Thm. 2) We will focus on the time $t = 3$. The proof for the general case follows naturally. Consider Fig. 1 which describes MDPUC at time $t = 3$, we want to show that:

$$P\left(s^{(3)}_{x([1,2])} \mid s^{(2)}_{x(1)}, x'^{(2)}_{x(1)}, s^{(1)}, x'^{(1)}\right) = P\left(s^{(3)}_{x(2)} \mid s^{(2)}, x'^{(2)}\right), \quad (7)$$

$$P\left(y^{(2)}_{x([1,2])} \mid s^{(2)}_{x(1)}, x'^{(2)}_{x(1)}, s^{(1)}, x'^{(1)}\right) = P\left(y^{(2)}_{x(2)} \mid s^{(2)}, x'^{(2)}\right). \quad (8)$$

Since Fig. 1 consists of 3 C-components: $S^{(1)}, \{X^{(1)}, Y^{(1)}, S^{(2)}\}$ and $\{X^{(2)}, Y^{(2)}, S^{(3)}\}$, Lem. 1 implies

$$\left(s^{(3)}_{x^{(2)}, s^{(2)}}, x'^{(2)}_{s^{(2)}} \perp\!\!\!\perp s^{(1)}, x'^{(1)}_{s^{(1)}}, s^{(2)}_{x^{(1)}, s^{(1)}}\right), \quad (9)$$

where $s^{([1,3])}, y^{([1,2])}, x^{([1,2])}, x'^{([1,2])}$ are arbitrary values. By composition and weak union axioms, we have

$$\left(s^{(3)}_{x^{(2)}, s^{(2)}} \perp\!\!\!\perp s^{(1)}, x'^{(1)} \mid s^{(2)}_{x^{(1)}}, x'^{(2)}_{x^{(1)}}\right). \quad (10)$$

Let $x'^{(1)} = x^{(1)}$ and apply composition and weak union axioms again.

$$\left(s^{(3)}_{x^{(2)}, s^{(2)}} \perp\!\!\!\perp s^{(1)}, x^{(1)} \mid s^{(2)}, x'^{(2)}\right). \quad (11)$$

Eqn. 11 implies

$$P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}, x'^{(2)}\right) = P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}, x'^{(2)}, s^{(1)}, x^{(1)}\right).$$

By Composition axiom, we move $x^{(1)}$ to subscripts.

$$\begin{aligned} &P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}, x'^{(2)}, s^{(1)}, x^{(1)}\right) \\ &= P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}_{x^{(1)}}, x'^{(2)}_{x^{(1)}}, s^{(1)}, x^{(1)}\right) \\ &= P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}_{x^{(1)}}, x'^{(2)}_{x^{(1)}}\right). \end{aligned}$$

The last step holds by the independence implied by Eqn. 10. Since $x^{(1)}$ in Eqn. 10 can be any value, let $x^{(1)} = x'^{(1)}$ and apply Eqn. 10 again.

$$\begin{aligned} &P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}_{x^{(1)}}, x'^{(2)}_{x^{(1)}}\right) \\ &= P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}_{x^{(1)}}, x'^{(2)}_{x^{(1)}}, s^{(1)}, x'^{(1)}\right) \\ &= P\left(s^{(3)}_{x^{([1,2])}, s^{(2)}} \mid s^{(2)}_{x^{(1)}}, x'^{(2)}_{x^{(1)}}, s^{(1)}, x'^{(1)}\right). \end{aligned}$$

The last step holds, since $s^{(3)}_{x^{(2)}, s^{(2)}} = s^{(3)}_{x^{([1,2])}, s^{(2)}}$ (Exclusion Restrictions rule). Together, we have

$$\begin{aligned} &P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}, x'^{(2)}\right) \\ &= P\left(s^{(3)}_{x^{([1,2])}, s^{(2)}} \mid s^{(2)}_{x^{(1)}}, x'^{(2)}_{x^{(1)}}, s^{(1)}, x'^{(1)}\right). \end{aligned} \quad (12)$$

We can rewrite $P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}, x'^{(2)}\right)$ as

$$\sum_{x'^{(1)} \in X, s'^{(1)} \in S} \underbrace{P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}, x'^{(2)}, s'^{(1)}, x'^{(1)}\right)}_{Term1} P\left(s'^{(1)}, x'^{(1)} \mid s^{(2)}, x'^{(2)}\right). \quad (13)$$

Some algebra through Composition axiom and Exclusion Restrictions rule turns Term 1 into:

$$\begin{aligned} &P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}, x'^{(2)}, s'^{(1)}, x'^{(1)}\right) \\ &= P\left(s^{(3)}_{x^{(2)}} \mid s^{(2)}, x'^{(2)}, s'^{(1)}, x'^{(1)}\right). \end{aligned} \quad (14)$$

Replace Term 1 with Eqn. 14, Eqn. 13 equals to:

$$\begin{aligned} &P\left(s^{(3)}_{x^{(2)}, s^{(2)}} \mid s^{(2)}, x'^{(2)}\right) \\ &= \sum_{x'^{(1)} \in X, s'^{(1)} \in S} P\left(s^{(3)}_{x^{(2)}}, s'^{(1)}, x'^{(1)} \mid s^{(2)}, x'^{(2)}\right) \\ &= P\left(s^{(3)}_{x^{(2)}} \mid s^{(2)}, x'^{(2)}\right) \end{aligned}$$

Together with Eqn. 12, we prove Eqn. 7. Eqn. 8 can be proved with the same steps but replacing $s^{(3)}$ with $y^{(2)}$. □

Thm. 2 proves the Markov property for a HA agent – in words, it says that the history $h'^{(t)}$ is best summarized by the current state $s^{(t)}$ and the physician's decision $x'^{(t)}$. We can then construct a MDP $M' = \langle S', X, T, R \rangle$ with an augmented state variable such that $S' = S \times X$, T and R are respectively Eqn. 7 and 8. To solve for the optimal policy for the HA agent in MDPUC is equivalent to solve for the optimal policy in M'.

When the full parametrization of MDPUC is provided, Eqn. 7 and 8 can be calculated through three-step procedure of Abduction, Action and Prediction [8, Sec. 7.1]. In practice, however, it is often difficult to obtain the distribution of unobserved variables ($P(U^{(t)})$). In such cases, Eqn. 7 and 8 can still be obtained through the new randomization procedure introduced in [1]: 1. observe the current state $s^{(t)}$ and the physician's decision $x'^{(t)}$; 2. stop the action $x'^{(t)}$ and perform a new action $x^{(t)}$ at random; 3. observe outcomes ($s^{(t+1)}$ and $y^{(t)}$), and record the data

$(s^{(t)}, x'^{(t)}, x^{(t)}, s^{(t+1)}, y^{(t)})$. Furthermore, the counterfactual representations of Eqn. 7 and 8 also suggest a off-policy learning method that can be useful to speedup convergence.

COROLLARY 1. *The MDPUC satisfies following statements for $t \geq 1$:*

$$P\left(s^{(t+1)}_{x^{(t)}}, x^{(t+1)}_{x^{(t)}} \mid s^{(t)}, x^{(t)}\right) = P\left(s^{(t+1)}, x^{(t+1)} \mid s^{(t)}, x^{(t)}\right)$$

$$P\left(y^{((t)}_{x^{(t)}} \mid s^{(t)}, x^{(t)}\right) = P\left(y^{(t)} \mid s^{(t)}, x^{(t)}\right)$$

PROOF. Since we condition on $x^{(t)}$, the simple application of the composition axiom [8, pp. 229] removes subscripts on left-hand side terms for both equations, which are exactly terms on the right-hand side. □

Note that quantities on the right hand side are observational distributions (without subscripts), which can be obtained by simply observing physicians naturally acting. In fact, when $x'^{(t)} = x^{(t)}$, Eqn. 7 and 8 can be estimated without direct experiments, but through the observational data alone.

Some reader might surmise that the results in Thm. 2 are immediate by arguing that the physician's decision $x'^{(t)}$ can be modeled as an extra state variable $X'^{(t)}$. The new state $\{S^{(t)}, X'^{(t)}\}$ must be Markov, since $X'^{(t)}$ only depends on independent local variables except for $S^{(t)}$. It is, therefore, not necessary, one may conclude, to perform any independence analysis among counterfactual variables. This statement, however, is certainly not true. Consider the dynamic system shown in Fig. 1(b), which is the same as Fig. 1(a) with the UC $U^{(t)}$ affecting the current state $S^{(t)}$. Even though the physician's decision $X^{(t)} = x'^{(t)}$ is still only affected by $U^{(t)}$ and $S^{(t)}$, the system is no long Markov for a HA agent. To witness, the confounding between $X^{(t)}$ and $S^{(t)}$ violates the independence relation in Eq. 9, which breaks the Markovian property. This means that the SCM in Fig. 1(b) cannot be cast to a MDP for a HA agent. This example illustrates that the independence relationships among counterfactual variables need to be carefully considered when solving sequential decision problems in natural settings where UCs exist. If the system is not Markovian, we have to resort to a more general formalism that permits such a violation.

5. HUMAN-ASSISTED AGENT AS A POMDP

The Markov property of the sequential decision process can be violated as the uncertainty of state information grows, which calls for the formalism of POMDPs. Fig. 3(a) shows the graphical representation of a prototypical POMDP model. POMDPs represent one of the most general formalisms for sequential decision problems. For example, the system in Fig. 3(a) can be translated into a POMDP by defining the state $S'^{(t)}$ in the POMDP as $S'^{(t)} = \{S^{(t)}, U^{(t)}\}$ and its observation $O^{(t)} = \{S^{(t)}\}$. For simplicity, we denote the POMDP state $S'^{(t)}$ by $S^{(t)}$. We focus on the planning problem for the POMDP. In a POMDP, the outcome of an action is related to all history since the POMDP is no longer Markovian. The solution, proposed in [4], is to introduce a belief state $B^{(t)}$ storing a probability distribution over the state space S up to time t given the history $H^{(t)} = h^{(t)}$. Formally, let $h^{(t)} = \{o^{(1)}, o^{(2)}_{x^{(1)}}, \ldots, o^{(t)}_{x^{([1,t-1])}}\}$,

$$B^{(t)}(s^{(t)}) = P\left(s^{(t)}_{x^{([1,t-1])}} \mid h^{(t)}\right) \tag{15}$$

Fig. 3(b) shows the graphical representation of a POMDP with a standard POMDP agent deployed. At time t, the agent receives the observation $o^{(t)}$, updates the belief state $B^{(t)}$, and picks an action based on the belief state. Given the current belief $B^{(t)}(s^{(t)})$, the most recent action $x^{(t)}$ and the most recent observation $o^{(t+1)}$, the belief state $B^{(t+1)}(s^{(t+1)})$ is updated by Bayes' rule:

$$\alpha P\left(o^{(t+1)}_{x^{(t)}} \mid s^{(t+1)}\right) \sum_{s^{(t)} \in S} P\left(s^{(t+1)}_{x^{(t)}} \mid s^{(2)}\right) B^{(t)}(s^{(t)}) \tag{16}$$

where α is a normalizing constant.

We next introduce the POMDP formalism for a HA agent. Following the same idea, we next introduce a belief state $B'^{(t)}$ storing the belief probability distribution over the state space S up to time t.

THEOREM 3. *Given the SCM described in Fig. 3(a), define the belief state*

$$B'^{(t)}(s^{(t)}) = P\left(s^{(t)}_{x^{([1,t-1])}} \mid h'^{(t)}\right),$$

where $h'^{(t)} = \{o^{(1)}, x'^{(t)}, o^{(2)}_{x^{(1)}}, x'^{(2)}_{x^{(1)}}, \ldots, o^{(t)}_{x^{([1,t-1])}}, x'^{(t)}_{x^{([1,t-1])}}\}$. Given the current belief $B'^{(t)}(s^{(t)})$, the most recent action $x^{(t)}$, the most recent physician's decision $x'^{(t+1)}$, and the most recent observation $o^{(t+1)}$, the next belief state $B'^{(t+1)}(s^{(t+1)})$ can be updated by Bayes' rule:

$$\alpha P\left(o^{(t+1)}_{x^{(t)}}, x'^{(t+1)}_{x^{(t)}} \mid s^{(t+1)}\right) \sum_{s^{(t)} \in S} P\left(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}\right) B'^{(t)}(s^{(t)}).$$

where α is a normalizing constant.

PROOF. By Bayes' rule,

$$B'^{(t+1)}(s^{(t+1)}) = P\left(s^{(t+1)}_{x^{([1,t])}} \mid o^{(t+1)}_{x^{([1,t])}}, x'^{(t+1)}_{x^{([1,t])}}, h'^{(t)}\right)$$

$$= \alpha P\left(s^{(t+1)}_{x^{([1,t])}}, o^{(t+1)}_{x^{([1,t])}}, x'^{(t+1)}_{x^{([1,t])}} \mid h'^{(t)}\right)$$

$$= \alpha P\left(o^{(t+1)}_{x^{([1,t])}}, x'^{(t+1)}_{x^{([1,t])}} \mid s^{(t+1)}_{x^{([1,t])}}, h'^{(t)}\right) P\left(s^{(t+1)}_{x^{([1,t])}} \mid h'^{(t)}\right).$$

A POMDP is Markov if $S^{(t)}$ is observed. By the Markov property, this turns into:

$$\alpha P\left(o^{(t+1)}_{x^{(t)}}, x'^{(t+1)}_{x^{(t)}} \mid s^{(t+1)}\right) P\left(s^{(t+1)}_{x^{([1,t])}} \mid h'^{(t)}\right). \tag{17}$$

By expanding on $s^{(t)}_{x^{([1,t-1])}}$ and the Markov property, $P\left(s^{(t+1)}_{x^{([1,t])}} \mid h'^{(t)}\right)$ is equivalent to

$$\sum_{s^{(t)} \in S} P\left(s^{(t+1)}_{x^{([1,t])}} \mid s^{(t)}_{x^{([1,t-1])}}, h'^{(t)}\right) P\left(s^{(t)}_{x^{([1,t-1])}} \mid h'^{(t)}\right)$$

$$= \sum_{s^{(t)} \in S} P\left(s^{(t+1)}_{x^{(t)}} \mid s^{(t)}\right) P\left(s^{(t)}_{x^{([1,t-1])}} \mid h'^{(t)}\right). \tag{18}$$

Note that $P\left(s^{(t)}_{x^{([1,t-1])}} \mid h'^{(t)}\right)$ is the current belief $B'^{(t)}(s^{(t)})$. Together with Eqn. 17 and 18, the theorem follows. □

Thm. 3 describes the belief update algorithm for a HA agent in a POMDP environment. Comparing the above formula

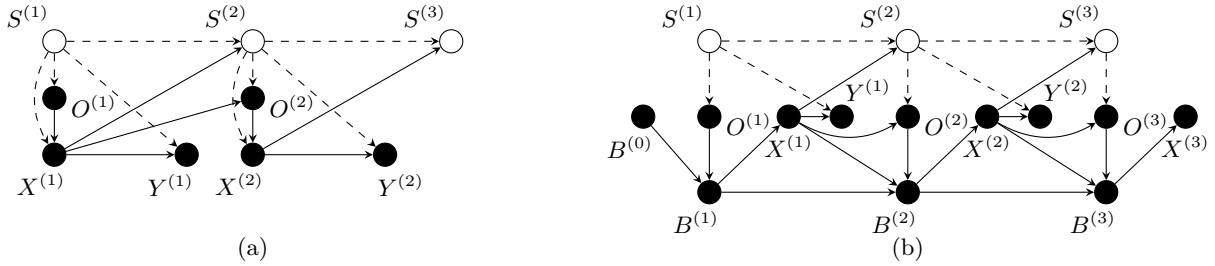


Figure 3: (a) The SCM representation for a POMDP where the human decides for $X^{(t)}$ based on $S^{(t)}, O^{(t)}$. (2) The SCM representation for a POMDP after a standard POMDP agent is deployed.

with Eqn. 16, we find that the physician's decision $x'^{(t)}$ acts as if it was an observations. This says that we can solve for the optimal policy for a POMDP HA agent by constructing a POMDP with a new observation $o'^{(t+1)} = \{o^{(t+1)}, x'^{(t+1)}\}$. The distribution $P\left(o^{(t+1)}_{x^{(t)}}, x'^{(t+1)}_{x^{(t)}} \mid s^{(t+1)}\right)$ can be calculated as

$$P\left(x'^{(t+1)} \mid s^{(t+1)}, o^{(t+1)}\right) P\left(o^{(t+1)}_{x^{(t)}} \mid s^{(t)}\right)$$

Let Π^*_{HA}, Π^*_{HR} denote the optimal policy of, respectively, the HA agent and the HR agent. Since $h^{(t)} \subset h'^{(t)}$, Π^*_{HA} should always dominate Π^*_{HR} for an arbitrary sequential system starting from state $s^{(1)}$. We now study the condition when the HA agent does not outperform the HR agent.

THEOREM 4. *For a POMDP described in Fig. 3(a) with a policy* $\Pi = \pi^{([1,t])}$ *where* $\pi^{(t)} := H^{(t)} \times X \to [0,1]$ *starting from state* $s^{(1)}$*, the following statement holds:*

$$V^{\Pi^*_{HA}}(s^{(1)}) = V^{\Pi^*_{HR}}(s^{(1)}) \tag{19}$$

PROOF. Since Π^*_{HA} always dominates Π^*_{HR}, it suffices to show the other direction $V^{\Pi^*_{HA}}(s^{(1)}) \leq V^{\Pi^*_{HR}}(s^{(1)})$. Suppose $\Pi^*_{HA} = \pi^{*([1,t])}_{HA}, \Pi^*_{HR} = \pi^{*([1,t])}_{HR}$, we write $V^{\Pi^*_{HA}}(s^{(1)})$ as:

$$\sum_{h^{(t)} \in H^{(t)}} \sum_{x^{([1,t])} \in X^t} \sum_{y^{(t)} \in Y} y^{(t)} P\left(y^{(t)}_{s^{(t)}, x^{(t)}}\right)$$
$$P\left(o^{(1)}_{s^{(1)}}\right) \prod_{i=2}^{t} P\left(o^{(i)}_{x^{(i-1)}, s^{(i)}}\right) P\left(s^{(i)}_{s^{(i-1)}, x^{(i-1)}}\right)$$
$$\prod_{j=1}^{t} \pi^{*(j)}_{HR}\left(x^{(j)} \mid h^{(j)}\right)$$

Similarly, we can write $V^{\Pi^*_{HR}}(s^{(1)})$ as:

$$\sum_{h'^{(t)} \in H'^{(t)}} \sum_{x^{([1,t])} \in X^t} \sum_{y^{(t)} \in Y} y^{(t)} P\left(y^{(t)}_{s^{(t)}, x^{(t)}}\right)$$
$$P\left(o^{(1)}_{s^{(1)}}\right) \prod_{i=2}^{t} P\left(o^{(i)}_{x^{(i-1)}, s^{(i)}}\right) P\left(s^{(i)}_{s^{(i-1)}, x^{(i-1)}}\right)$$
$$\prod_{j=1}^{t} P\left(x'^{(j)} \mid h^{(j)}\right) \pi^{*(j)}_{HA}\left(x^{(j)} \mid h'^{(j)}\right)$$

Since $h'^{(t)} = h^{(t)} \cup \{x'^{(1)}, \ldots, x'^{(t)}_{x^{([1,t-1])}}\}$, this turns to

$$\sum_{h^{(t)} \in H^{(t)}} \sum_{x^{([1,t])} \in X^t} \sum_{y^{(t)} \in Y} y^{(t)} P\left(y^{(t)}_{s^{(t)}, x^{(t)}}\right)$$
$$P\left(o^{(1)}_{s^{(1)}}\right) \prod_{i=2}^{t} P\left(o^{(i)}_{x^{(i-1)}, s^{(i)}}\right) P\left(s^{(i)}_{s^{(i-1)}, x^{(i-1)}}\right)$$
$$\underbrace{\sum_{x'^{([1,i])} \in X^i} \prod_{j=1}^{t} P\left(x'^{(j)} \mid h^{(j)}\right) \pi^{*(j)}_{HA}\left(x^{(j)} \mid h'^{(j)}\right)}_{Term1}$$

Term 1 defines a policy of the HR agent after summing out $x'^{([1,t])}$, which we denote by Π_{HR}. Since Π^*_{HR} is the optimal policy for the HR agent, it follows that

$$V^{\Pi^*_{HA}}(s^{(1)}) = V^{\Pi_{HR}}(s^{(1)}) \leq V^{\Pi^*_{HR}}(s^{(1)}).$$

□

Thm. 4 concerns with the value of information of the human's decision. When the state spaces of the human and the agent coincide, i.e., the human does not observe more information than the agent, the optimal performance of a HR agent matches a HA agent. Therefore, it is completely safe under these conditions to replace the human with an automated agent. In general, however, even when biased (and possibly worse than a random policy), the information coming from the human decision-maker allows an HA agent to dominate any traditional HR agent.

6. APPLICATIONS AND EXPERIMENTS

Our goal in this section is to operationalize the learning algorithms for the HA agent in both offline planning and online learning settings. We focus on the SCM of the medical treatment example described in Sec. 3. The physician's policy is defined as

$$X^{(t)} \leftarrow \pi^{ndt}(s^{(t)}, m^{(t)}, s^{(t)}) = s^{(t)} \oplus m^{(t)} \oplus e^{(t)},$$

where $\oplus$ represents the exclusive OR operator.

The reward probability function $P(y^{(t)} \mid s^{(t)}, m^{(t)}, e^{(t)}, x^{(t)})$ and the transition probability function $P(s^{(t+1)} \mid s^{(t)}, x^{(t)})$ are provided in Tables 1 and 2. The entries encode the probabilities for $Y^{(t)} = 1$. The doctor's natural choice of action (i.e., following π^{ndt}) are indicated by asterisks.

Evaluation Metrics. The performance is evaluated with standard metrics: (1) the cumulative reward per episode

	$S^{(t)}=0$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	$*0.2$	0.9	0.8	$*0.3$
$X^{(t)}=1$	0.9	$*0.2$	$*0.3$	0.8
	$S^{(t)}=1$			
	$M^{(t)}=0$		$M^{(t)}=1$	
	$E^{(t)}=0$	$E^{(t)}=1$	$E^{(t)}=0$	$E^{(t)}=1$
$X^{(t)}=0$	0.7	$*0.2$	$*0.1$	0.8
$X^{(t)}=1$	$*0.2$	0.7	0.8	$*0.1$

Table 1: Reward probability table for health score $Y^{(t)}=1$, which is $P(Y^{(t)}=1 \mid s^{(t)}, m^{(t)}, e^{(t)}, x^{(t)})$. The doctor's natural choice under $S^{(t)}, M^{(t)}, E^{(t)}$ are indicated by asterisks.

	$S^{(t)}=0$	$S^{(t)}=1$
$X^{(t)}=0$	0.9	0.3
$X^{(t)}=1$	0.7	0.8

Table 2: The transition probability table $P(S^{(t+1)}=0 \mid s^{(t)}, x^{(t)})$.

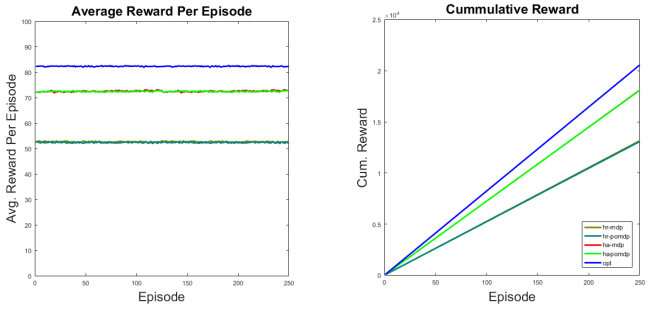
averaging over 800 runs (AR), and (2) the cumulative reward for 250 episodes (CR).

Experiment 1: Offline Planning We run experiments for the offline planning setting of the medical treatment example, where full parametrizations of the model are provided to the agent. We compare both the MDP and POMDP planning algorithms for HR and HA agents: MDP value iteration for a HR agent (labeled *hr-mdp*), MDP value iteration for a HA agent (*ha-mdp*), POMDP incremental pruning for a HR agent (*hr-pomdp*), POMDP incremental pruning for a HA agent (*ha-pomdp*). We also include the optimal policy computed by the agent who can access all unobserved states (*opt*). We believe this is fair for our examples since it allows the comparison of our algorithm against a truly optimal policy with full access to the unobserved variables.

Results shown in Fig. 4 support the HA agent approach. The simulation reveals the HA agent (CR=2.0211×10^4) consistently outperforms the HR agent (CR=1.3076×10^4) in both MDP and POMDP planning algorithms. Also, we note that the POMDP planning algorithm learns the same policy as the one learned by the MDP algorithm for the HA agent. This confirms that the system consisting of the MDPUC model and a HA agent is Markov and can be solved by MDP algorithms (Thm. 2).

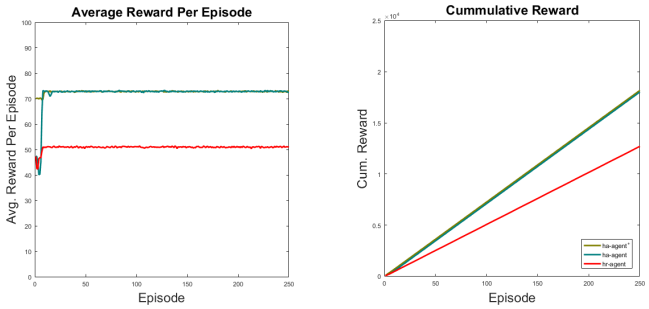
Experiment 2: Online Learning We run experiments for the online learning setting of the medical treatment example, where the parametrizations of model are not available to the agent. The agent has to converge to the optimal policy while explores the environment. We compare three variants of the MORMAX [12] algorithm: MORMAX for a HR agent (labeled *hr-agent*), MORMAX for a HA agent (*ha-agent*), and MORMAX for a HA agent with samples transfer method in Col. 1 (*ha-agent*$^+$).

Results shown in Fig. 5 support the HA agent approach. Specifically, the simulation reveals an improvement in cumulative reward obtained by *ha-agent*$^+$ (2.0574×10^4) compared to *ha-agent* (2.0335×10^4). We can also see from the average reward per episode that *ha-agent*$^+$ shows a faster convergence rate than *ha-agent*, which corroborates with the view that the sample transferring procedure (Corol. 1) leveraging



(a) (b)

Figure 4: Simulation results for Experiment 1 comparing the offline planning performance for HR and HA agents with both MDP and POMDP planning methods.



(a) (b)

Figure 5: Simulation results for Experiment 2 comparing the online learning performance for a HR agent, a HA agent and a HA agent while leveraging observational data through seeding.

observational data can be helpful. Surprisingly, *ha-agent*$^+$ is able to converge to an optimal policy in the very beginning. The *hr-agent*, predictably, is not a competitor and experiences a relatively low cumulative reward (1.3278×10^4).

Overall, these results confirm that the HA agent, which utilizes the the natural agent's decision, converge to a higher expected return; the samples transferring procedure allows algorithms to converge at a faster pace in the online settings.

7. CONCLUSION

We studied the problem of finding optimal decision-making strategies when a natural agent is already deployed and decision are possibly driven by unobserved confounders. Using counterfactual machinery, we delineated two classes of agents, namely, HR and HA – the former class attempts to completely replace humans agents, while the latter attempts to collaborate with them to reach better decisions. We first showed that an optimal strategy could be found by modeling the HA agent as a modified version of MDP and POMDP solvers, depending on whether the assumption of counterfactual Markovianity holds in the environment. Through a syntactic transformation of the state variable, we operationalized these strategies and showed that HA agents consistently dominate their HR counterparts. Finally, we derived the conditions when the performance of both agents coincide, which delineates the class of problems where human input does not contain useful information.

REFERENCES

[1] E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2015.

[2] D. P. Bertsekas and J. N. Tsitsiklis. Neuro-dynamic programming: an overview. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1, pages 560–564. IEEE, 1995.

[3] A. Cassandra, M. L. Littman, and N. L. Zhang. Incremental pruning: A simple, fast, exact method for partially observable markov decision processes. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 54–61. Morgan Kaufmann Publishers Inc., 1997.

[4] A. R. Cassandra, L. P. Kaelbling, and M. L. Littman. Acting optimally in partially observable stochastic domains. In *AAAI*, volume 94, pages 1023–1028, 1994.

[5] M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011.

[6] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.

[7] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

[8] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

[9] D. Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[10] S. Ross, J. Pineau, S. Paquet, and B. Chaib-Draa. Online planning algorithms for pomdps. *Journal of Artificial Intelligence Research*, 32:663–704, 2008.

[11] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[12] I. Szita and C. Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1031–1038, 2010.

[13] J. Tian and J. Pearl. On the identification of causal effects. Technical Report R-290-L, Department of Computer Science, University of California, Los Angeles, CA, 2003. http://www.cs.iastate.edu/∼jtian/r290-L.pdf.